УДК 004.912+004.85

МОДИФИКАЦИИ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ РЕШЕНИЯ ЗАДАЧИ ИДЕНТИФИКАЦИИ АВТОРА ТЕКСТА

канд. техн. наук, доц. А. И. ПАРАМОНОВ,
магистр техн. наук И. А. ТРУХАНОВИЧ
(Белорусский государственный университет информатики и радиоэлектроники, Минск)

Аннотация. *В процессе решения задачи идентификации автора текста хорошо зарекомендованы методы машинного обучения. В работе рассмотрены некоторые направления модификации методов машинного обучения для повышения точности идентификации автора. Описаны результаты проведенных экспериментов.*

Ключевые слова: *идентификация авторства, машинное обучение, рекуррентные нейронные сети.*

Проблема автоматической обработки массивов неструктурированных текстов с целью категоризации цифровых документов сегодня все чаще возникает в разных отраслях деятельности. При этом классифицировать тексты предполагается по различным критериям в зависимости от поставленной задачи. Одними из значимых задач обработки естественно-языковых текстов остаются: определение контекста, анализ эмотивности и идентификация авторства. В связи с постоянным ростом количества анонимных цифровых текстов, в том числе к глобальной сети, все более остро и актуально поднимается вопрос установления автора неизвестного текста.

Для идентификации авторства часто используется экспертный анализ. Эксперты могут идентифицировать авторов по различным языковым признакам. Однако работа экспертов занимает значительное время, что затрудняет обработку больших массивов данных. Автоматизация этого процесса помогает не только при работе с большим количеством текстов, но и в целом упрощает идентификацию авторства. Существующие программные подходы позволяют учитывать и варьировать различные параметры, характеризующие текст [1].

Автоматизация решений задачи идентификации автора текста предполагает решения ряда подзадач, а именно: выбор модели представления текста, выделение набора признаков автора и подбор метода классификации. В качестве методов классификации могут быть использованы количественные методы и методы статистического анализа. Однако большая эффективность сегодня достигается за счет применения методов машинного обучения [2–3].

Среди основных методов машинного обучения наиболее популярен подход с использованием нейронных сетей. Нейронные сети могут использовать векторный слой, на котором текст преобразуется в закодированное представление для последующей обработки остальной частью сети. Тексты могут быть представлены, например, с помощью мешка слов. Установлено, что выбор слоев сети зависит от того, какая у нейронной сети будет общая архитектура. В рамках исследования были рассмотрены некоторые известные методы машинного обучения и варианты их модификаций для применения в решении задачи идентификации автора текста.

В качестве представления текста предлагается использовать вектор признаков. Одна из наиболее известных проблем этого подхода заключается в том, что признаки, которые подходят для одной группы текстов, могут не подходить для другой. Данная проблема может быть решена с помощью генетического алгоритма. Назначив поиск эффективного набора признаков для определенных текстов в качестве цели для генетического алгоритма, можно добиться значительного прогресса в результатах. В качестве классификатора могут быть выбраны методы случайного леса, логистическая регрессия и другие.

Следующим шагом экспериментов после выбора модели представления текста будет выбор архитектуры сети. При обработке текстов сегодня широко известен и применяется подход рекуррентных нейронных сетей, поскольку они дают хорошие результаты при обработке последовательностей. Среди используемых модификаций рекуррентных сетей выделяются долгая краткосрочная память и управляемый рекуррентный блок.

Долгая краткосрочная память – это тип рекуррентной нейронной сети, специально разработанный для того, чтобы предотвратить затухание или рост выходного сигнала нейронной сети для заданного входного сигнала по мере прохождения через циклы обратной связи. Именно циклы обратной связи позволяют рекуррентным сетям быть лучше в распознавании, чем другие нейронные сети. Память на прошлые входные данные критически важна для решения задач обучения последовательности, и сети с долгой краткосрочной памятью обеспечивают лучшую производительность по сравнению с другими рекуррентными нейронными сетями за счет работы с так называемой проблемой исчезающего градиента [4].

Управляемый рекуррентный блок – это тип рекуррентной нейронной сети, которая в некоторых случаях имеет преимущества перед долгой краткосрочной памятью. Управляемый рекуррентный блок решают проблему исчезающего градиента с помощью двух вентилей - вентилей обновления и вентилей сброса. Эти вентили решают, какая информация будет пропущена на выход, и могут быть обучены удерживать информацию из более далекого прошлого. Это позволяет передавать соответствующую информацию вниз по цепочке событий для более точного прогнозирования [5].

Управляемый рекуррентный блок задан уравнениями (1).

$$r = \sigma(W_1 x_t + W_2 h_{t-1} + b_1),$$

$$z = \sigma(W_3 x_t + W_4 h_{t-1} + b_2), \quad (1)$$

$$m_t = tanh(W_5 x_t + W_6(r \odot h_{t-1}) + b_3),$$

$$h_t = z \odot m_t + h_{t-1} \odot (1 - z),$$

где W_1, W_2, W_3, W_4, W_5, W_6 – матрицы параметров;

b_1, b_2, b_3 – вектора смещения;

x_t, h_t – входной и выходной вектор;

Θ – поэлементное умножение векторов;

σ, tanh – функции активации (сигмоида и гиперболический тангенс).

Тем не менее, возможности модификации для рекуррентных сетей еще не исчерпаны и могут быть предложены новые варианты. Поиск этих вариантов также может быть выполнен с помощью генетического подхода, предполагая те или иные комбинации компонентов нейронной сети.

В рамках эксперимента были рассмотрены указанные подходы по конфигурации сети и на тестовом наборе проведено их сравнение по точности идентификации для тестовой выборки из набора (после обучения). Точность будет отображать процент правильного соотнесения неизвестных текстов с авторами. В качестве исходных текстов для идентификации авторов может быть использован набор Reuter_5050, который содержит по 50 текстов от каждого из 50 авторов [6].

Первая серия экспериментов был проведена на конфигурации с представлением текстов в виде векторов признаков и с применением классификаторов – случайный лес и линейный дискриминантный анализ. Предполагалось выявить наиболее подходящие наборы признаков, обеспечивающие максимальную точность для этих классификаторов.

Во второй серии экспериментов применена известная модификация долгой краткосрочной памяти (управляемый рекуррентный блок), а также предложена новая модификация, которая получена путем генетического поиска – задана формулами (2).

$$z = \sigma(W_1 x_t + b_1),$$

$$r = \sigma(W_2 x_t + W_3 h_t + b_2), \quad (2)$$

$$h_{t+1} = tanh(W_3(r \odot h_t) + tanh(x_t) + b_3) \odot z + h_t \odot (1 - z),$$

где W_1, W_2, W_3 – матрицы параметров;

b_1, b_2, b_3 – вектора смещения;

x_t, h_t – входной и выходной вектор;

Θ – поэлементное умножение векторов;

σ, tanh – функции активации (сигмоида и гиперболический тангенс).

Результаты эксперимента приведены в таблице 1. При этом, для случайного леса и линейного дискриминантного анализа были выбраны три наиболее точных набора признаков.

Таблица 1 – Результаты эксперимента

Метод	Точность
Случайный лес (первый набор признаков)	90.9
Случайный лес (второй набор признаков)	90.5
Случайный лес (третий набор признаков)	90.2
Линейный дискриминантный анализ (первый набор признаков)	61.9
Линейный дискриминантный анализ (второй набор признаков)	49.4
Линейный дискриминантный анализ (третий набор признаков)	46.9
Долгая краткосрочная память (модификация)	57.1
Управляемый рекуррентный блок	55.3

Полученные результаты дают предпосылки, что путем модификаций методов (как рассмотренных в рамках эксперимента, так и других) можно достичь прироста точности идентификации. Кроме того, есть возможности для комбинации различных модификаций.

ЛИТЕРАТУРА

1. Батура, Т. В. Формальные методы определения авторства текстов / Т. В. Батура. — Новосибирск : Вестник НГУ, серия «Информационные технологии», Том 10, Выпуск 4, 2012. – С. 81–94.
2. Парамонов, А. И. Методы анализа цифрового текста для идентификации его автора / А. И. Парамонов, И. А. Труханович // Веб-программирование и интернет-технологии WebConf2021 : материалы 5-й международной научно-практической конференции, Минск, 18–21 мая 2021 г. / Белорусский государственный университет ; редкол.: И. М. Галкин [и др.]. – Минск, 2021. – С. 118–119.
3. Paramonov, A. Dynamic features selection in authorship identification problem / A. Paramonov, I. Trukhanovich, U. Kuntsevich // Open Semantic Technologies for Intelligent Systems (OSTIS-2021) : сборник научных трудов / Белорусский государственный университет информатики и радиоэлектроники ; редкол. : В. В. Голенков [и др.]. – Минск, 2021. – Вып. 5. – С. 309–312.
4. Long Short-Term Memory (LSTM) [Электронный ресурс]. – Режим доступа: https://developer.nvidia.com/discover/lstm. – Дата доступа: 18.03.2022.
5. Gated Recurrent Unit [Электронный ресурс]. – Режим доступа: https://blog.marketmuse.com/glossary/gated-recurrent-unit-gru-definition/. – Дата доступа: 20.03.2022.
6. Reuter_50_50 [Электронный ресурс]. – Режим доступа: https://archive.ics.uci.edu/ml/datasets/Reuter_50_50. – Дата доступа: 21.03.2022.